SIGNIFICANCE

Physical methods are poised to transform drug discovery and chemical biology by enabling true molecular design. While modeling work is used extensively in drug discovery, its main role at present is to aid with idea generation or to filter large libraries of compounds for screening. Instead, we imagine using computational techniques extensively to guide the design process. Consider a medicinal chemist in the not-too-distant future who has just finished synthesizing several new derivatives of an existing inhibitor as potential drug leads targeting a particular biomolecule, and has obtained binding affinity or potency data against the desired biomolecular target. Before leaving work, he or she generates ideas for perhaps 100 new compounds which could be synthesized next, then sets a computer to work overnight prioritizing them. By morning, the compounds have all been prioritized based on reliable predictions of their affinity for the desired target, selectivity against alternative targets which should be avoided, solubility, and membrane permeability. The chemist then looks through the predicted properties for the top few compounds and selects the next ones for synthesis. If synthesizing and testing each compound takes several days, this workflow compresses roughly a year's work into a few days.

While this workflow is not yet a reality, huge strides have been made in this direction, with calculated binding affinity predictions now showing real promise [**?**, **?**, **?**, **?**, **?**, **?**, 1, 2], rapid progress toward solubility predictions [**?**, **?**, **?**], and selectivity and drug resistance also apparently tractable [**?**], with some headway apparent on membrane permeability [**?**, **?**]. A considerable amount of science and engineering still remains to make this vision a reality, but **given recent progress, the question now seems more one of *when* rather than *whether*.**

The widespread availability of inexpensive graphics processing units (GPUs), which provide a 100-fold increase in price/performance over CPUs, coupled with advances in automation [**?**] and sampling protocols have helped simulation-based techniques reach the point where they now begin to be genuinely useful in guiding drug discovery for a limited *domain of applicability* [**?**, **?**, **?**, **?**, **?**, **?**, **?**]. Specifically, in some situations, free energy calculations appear to be capable of achieving RMS errors of 1-2 kcal/mol with current force fields, even in prospective applications, sufficient to drastically reduce the number of molecules that must be synthesized and assayed [**?**]. As a consequence, pharmaceutical companies are beginning to use these methods in active discovery projects.

Despite this progress, these methods currently have severe limitations, and a great deal of science and engineering is still needed before these techniques can achieve their full potential in transforming molecular design. For example, even "small" protein conformational conformational changes not gracefully handled by current methodologies can yield errors up to 5 kcal/mol in calculated binding free energies [3], force field limitations still pose major challenges [**?**], and the inability to treat important chemical effects like protonation and tautomer equilibria drastically limits the domain of applicability. For many pharmaceutically relevant systems, the most important sources of error are not yet clear.

Progress on addressing these challenges has been frustratingly slow, hindered by a lack of high-quality data and community focus. **Neither retrospective tests nor prospective application in active drug discovery projects provides the necessary impetus and data to rapidly overcome the remaining major barriers to widespread utility.** Large-scale retrospective tests certainly assess our *retrospective* performance, but they do not provide accurate guidance on utility for prospective design, nor do they effectively pave the way forward. One issue with retrospective tests is that they can easily result in over-fit models, or researchers applying a variety of different protocols until they get good results, perhaps even by a statistical fluke [ref] . In retrospective tests, performance may also not be indicative of expected performance in applications because even well-meaning researchers can take advantage of prior knowledge. For example, if the binding mode of a ligand is already known crystallographically, a researcher may use that binding mode in retrospective tests, whereas prospective or design work would require first selecting among candidate binding modes, introducing substantial uncertainty unaccounted for in the retrospective statistics [1, 4, 5]. This also means that in retrospective tests, researchers almost invariably try far fewer methods than in prospective tests, resulting in much less new insight. Prospective tests, in contrast, force researchers to worry about a multitude of potential situations rather than only those observed in a known benchmark dataset. Prospective application in actual discovery projects, while important, also does not provide the necessary impetus, partly because often, the predicted compounds are in fact never tested [**?**] or the experimental data necessary to assess the quality of the predictions is absent—for example, because binding affinities are not measured or no crystallography is available.

To accelerate progress in quantitative predictive physical modeling, we need a series of community blind prediction challenges focused on pushing the limits of predictive techniques, beginning from problems which are just barely tractable with today's methods and advancing to problems just past the frontier. These challenges should be designed to have precisely the necessary high quality experimental data, but also be prospective, predictive tests. While the Drug Design Data Resource (D3R [6], discussed further below) provides an

existing community blind challenge on protein-ligand binding, it focuses on using *existing* pharmaceutical datasets for blind challenges, and not on introducing new data in a carefully controlled manner in order to maximize the learning value to the community [6]. In other words, D3R serves well to assess where we are now – but we need a carefully designed effort that will help the field achieve our goals.

In our experience, physical modeling advances most rapidly when confronted by specific problems which must be modeled accurately, as revealed by carefully collected and curated data. Thus, we need an effort which focuses on specific *component* problems of the overall problem of interest, collects and curates data that highlights these problems, and then brings this data to the community to drive progress via prospective challenges. This process allows the entire community to learn from what succeeds and fails in these challenges. The model we propose here has *already* worked to drive dramatic improvements in modeling, as evidenced by our **Statistical Assessment of Modeling of Proteins and Ligands (SAMPL)** series of challenges. SAMPL was born out of frustration with the lack of venues for comparing predictive accuracy on a level playing field and was initiated by Anthony Nicholls of OpenEye software in 2007/2008 [7], and has run approximately every two years since then [8–15]. Governance transitioned to an unfunded academic collaboration during SAMPL3 in 2012; this collaboration ran subsequent challenges as SAMPL4 (2014) and SAMPL5 (2016). The PI of this proposal (Mobley) played a key organizational role in SAMPL3–SAMPL5. SAMPL geared around the idea that learning collectively from failures and successes, rather than competition, will provide the greatest long-term rewards in terms of progress in the field. SAMPL has always involved a component focused on calculation of relatively straightforward physical properties such as hydration free energies, but also introduced host-guest systems as model binding systems for SAMPL3-5, supplementing protein-ligand challenges (on trypsin and HIV integrase) which appeared in SAMPL3-4. SAMPL has already been a tremendous resource for the community, resulting in roughly 100 publications (some are coming out as of this writing) which are typically cited 5-50 times or more each [refs].

Here, we expand SAMPL by **designing a new series of SAMPL challenges specifically to maximize learning value to the community**. Until now, this has been impossible, because SAMPL has been entirely unfunded; its very existence has required "donation" of data and time from various sources, hindering the ability to collect datasets tailored to our purpose. The incarnation of SAMPL proposed here will deliberately bridge the gap between calculations of simple physical properties that isolate forcefield inaccuracies from sampling challenges, like hydration—which can be calculated fairly accurately with today's methods [12]—and the D3R Grand Challenges on protein-ligand binding, which are a major source of consternation for the community so far [6,16–18]. Unless this gap is bridged, there is the very real possibility that modeling may simply continue to fall far short of expectations in pharmaceutical challenges like D3R for reasons which are unclear. The extension of SAMPL proposed here will allow us to form blind challenges designed to highlight major reasons for failure and drive progress towards resolving them.

Our major goal is to rapidly advance predictive modeling to where it can guide experimental work doing biomolecular design, and extension of the SAMPL challenges will do exactly that. This work will play a vital role in enhancing the work being done on *existing* data by D3R, helping prepare methods for application to the more challenging systems emerging from pharma in D3R's challenges.

INNOVATION

Blind predictive challenges—and SAMPL in particular—have already led to important new science in the form of method development, evaluation, robustness, and force field development. But they have not yet resulted in a dramatic improvement in predictive molecular design, in part because the **lack of funding has left us unable to design a systematic and progressive series of challenges that drives such improvements**. Previous SAMPL challenges have been valuable in isolation – but they have not moved the frontier because they were driven by data availability rather than purposeful design. Here, we propose the innovative step of designing SAMPL to advance physical modeling.

Several historical examples serve to highlight how SAMPL can foster innovation (though far more examples are available [CITE]). The first several SAMPL challenges on hydration free energies had rather hit-and-miss performance, highlighting pitfalls of existing methods and force fields which led to marked improvements in PB models [refs], recognition of some limitations of fixed-charge force fields [refs], repair of some of these force field deficiencies via additional polarization or introduction of off-site charges [refs], and helped motivate alternate implicit or hybrid solvent models [ref]. Together these advances led to a marked improvement in accuracy for calculations of hydration free energies between SAMPLX and SAMPLY (Figure []). Shifts in protonation state and tautomer proved particularly important in the recent SAMPL5 logD challenge [refs]. This challenge provided a tractable opportunity to isolate and explore these specific physical effects, which are so important in protein-ligand binding, while avoiding the full complexity of pharmaceutical binding studies. Host-guest binding studies have also

been particularly important [ref benchmark sets], highlighting the importance of salt effects [refs] and in some cases revealing more severe force field limitations than observed in hydration and distribution challenges [ref Gilson group stuff], pointing the way forward for improving predictive models of molecular interactions [ref].

This work is also innovative because of the uniqueness of SAMPL. While there are other predictive challenges in the area of biomolecular modeling, such as D3R [ref], the pKa cooperative [ref], CAPRI [ref] and CASP [ref], **no other effort focuses specifically on driving quantitatively predictive protein-ligand modeling**. The SAMPL expansion we propose here is unique because it will be *specifically designed* to drive improvements in modeling accuracy for biomolecular design, rather than simply serving to evaluate accuracy on targets of pharmaceutical interest. It also plays an enabling role for a wide variety of other science, rather than functioning as a stand-alone entity. SAMPL benefits the whole modeling community—for example, protein-ligand docking software has improved as a direct result of SAMPL hydration challenges [ref Coleman], and commercial software vendors have introduced new features or scientific improvements based on participation in SAMPL challenges [cite Klamt stuff]. In effect, **SAMPL serves as an engine to spur innovation by new soliciting novel approaches to complex problems from the community and evaluating their success on blinded data.** The proposed project will ensure that SAMPL not only continues but becomes even more valuable to the community.

This work also focuses on innovative experimental methods. Specifically, in Aim 3, we develop a new informatics platform to facilitate the rapid identification and development of useful protein-ligand model systems that are both experimentally tractable for high-throughput biophysical measurements and focus on specific challenges of interest. We employ a fully automated wetlab to screen potential model systems for expression, carry out high-accuracy biophysical measurements, and perform automated error analysis to carefully assess experimental uncertainty. The Chodera lab—responsible for Aim 3—has substantial expertise in the area of automation of biophysical measurements and uncertainty analysis [**?**], and releases data analysis tools and 3D printed parts developed to facilitate these experiments as open source. **This work is at the forefront of innovation in high-throughput, automated biophysical experiments to produce high quality data with well-characterized uncertainties.** Not only will the data itself be of prime importance to the community, but the techniques themselves will help future experiments.

In Aim 4, we will not only run annual iterations of SAMPL community challenges, but also perform our own reference calculations with the latest techniques, testing their accuracy and using these to assess the current state-of-the-art. Both the Mobley and Chodera labs are experts in development of free energy methods for application to physical properties (e.g., [a couple refs]) and binding (e.g., [a couple refs]), and the **reference calculations we perform in Aim 4 are particularly important for innovation as well**, serving several key roles: (1) Benchmarking our latest method developments against current “best practices” methods (by doing calculations via both approaches); (2) Facilitating learning, allowing others to experiment with how a change in method or force field impacts results; (3) Focusing the field on key issues by doing sensitivity analysis to whether conditions such as ionic strength, protonation state, tautomer choice, etc., impact computed values.

Innovation in Aim 4 extends beyond reference calculations to analysis of the challenge itself. When methods differ in performance, it is critical to understand whether the differences are statistically significant and important, and to provide an accurate accounting of the uncertainty in performance measures. Thus, careful and innovative analysis of challenge outcomes is particularly important in SAMPL [refs], in some cases driving experimentation with new performance metrics [refs]. Tracking down *why* models are failing and what particular problems still need attention is an in important and powerful aspect of SAMPL. Much of this is left to the participants, but as organizers we do a great deal of work spotting patterns and providing a venue for participants to explore these issues. In Aim 4, here, we also extend the scope of our reference calculations, providing further opportunity to drive innovation in this way. Additionally, we try to draw attention to and promote analysis of model uncertainty (as distinct from statistical uncertainty) in calculated values [refs], as understanding the confidence levels of predictions is particularly important for guiding molecular design.

APPROACH

Our approach to systematically advancing modeling for biomolecular design involves collecting carefully targeted experimental datasets for challenges focusing on physical property prediction, host-guest binding, and protein-ligand binding that isolate individual limitations in quantitative physical modeling to solicit and evaluate multiple solutions from the community. Aims 1–3 focus on tailoring and generating this experimental data, while Aim 4 focuses on fielding annual SAMPL challenges. Aims 1–3 bring together multiple laboratories and both theorists and experimentalists: graduate students from the Mobley and Chodera laboratories are paired with well-equipped experimental groups in industry to collect physical property data (Aim 1); leading experimentalists in supramolecular chemistry Gibb and Isaacs work with theorist Mobley to perform host-guest affinity measurements (Aim 2); new

approaches to the automated development and investigation of model protein-ligand systems are developed in the Chodera lab (Aim 3). The annual SAMPL challenges organized by the Mobley and Chodera labs (Aim 4) will evaluate best-practices reference calculations, will feature both annual physical meetings (coordinated with D3R) and virtual communication platforms to maximize community engagement, develop novel statistical analyses to identify and respond to problem areas, release benchmark datasets alongside primary experimental data, organize special journal issues to report progress, and encourage tool automation whenever possible.

Aim 1: Isolate forcefield accuracy and chemical effects from sampling challenges by collecting new physical property measurement benchmark datasets. Our first aim focuses on generating solution-phase physical property data for small, drug-like molecules—essentially continuing the tradition begun with hydration free energies in SAMPL0-4 and continued with water-cyclohexane distribution coefficients in SAMPL5. Distribution coefficients proved tremendously valuable to the community in SAMPL5 for a number of reasons, highlighting a number of key issues where modeling needs to improve, so they will form the basis for the physical property component in SAMPL6 and return in several subsequent SAMPL challenges. [JDC: This first paragraph serves no purpose. The fact that we are continuing a tradition is the *least* most important point, so why lead with that? And instead of mentioning *why* these methods are tremendously valuable, you say "because reasons". Most likely, our reviewers will not easily be able to make the connection on their own that physical property measurements allow us to isolate deficiencies in our current physical modeling approaches—forcefield accuracy and chemical effects like protonation and tautomer states—from the more complex challenges in protein-ligand binding, such as very slow conformational changes that hinder convergence of calculations. Let's try to make this connection as easy as possible for them by explaining that from the outset. We want to feed them the phrase to write in the Summary Statement that says why this Aim is important and well-designed.]

Rationale: Distribution coefficients proved to be precisely the right level of difficulty for a SAMPL challenge focused on maximizing learning. [JDC: I don't like the phase "maximizing learning". Who is learning? What are they learning and how? Can we say something more specific? Distribution coefficients measure the transfer free energies from aqueous high-dielectric to nonpolar low-dielectric environments, which mimics many of the physical characteristics of protein binding sites but is free of the sampling challenges of slow conformational change or engaging specific polar or charge interactions. Despite this, phenomena of protonation and tautomeric state changes upon transfer persist, and forcefield deficiencies are brought to the forefront, without concern that the calculations are poorly converged. Can we communciate this?] Specifically, they were challenging enough that many methods performed poorly, with even the best methods having accuracies less than would have been expected based on their ability to calculate hydration free energies in water [19]. At the same time, methods typically did well enough that it was possible to learn a great deal from examining failure, and the major sources of error were issues which will also plague prediction of ligand-receptor interactions. These included neglect of changes in protonation state on transfer between environments, uncertainty as to the relevant protonation state and/or tautomer in one or both environments [19], problems with sampling the conformation of some of the larger ligands [19, 20], and force field limitations [21]. Our partnership with Genentech for these measurements also meant that the compounds were from Genentech's library and thus very drug-like, unlike typical compounds seen in hydration free energy challenges of the past. [JDC: That's not quite accurate. The compounds are commercially available and were selected to be drug-like and span a large log D range. I think the partnership with Genentech should be mentioned at the beginning of this paragraph in a two-sentence summary of the challenge and its goals. We also need to explain why we chose cyclohexane and not octanol for this initial challenge.] In some respects, distribution coefficients posed the ideal SAMPL challenge, hitting the sweet spot in terms of difficulty—difficult enough that clear failures were frequent and there is much room for improvement, but not so difficult that the reasons for failure were unclear in general. Still, the challenge could have been improved by follow-up experiments to re-check some of the experimental results [14, 21–23]—but without funding for someone to continue working in this space, this has so far been impossible. [JDC: Turn this into a positive statement: This exercise the utility that funded, targeted follow-up experiments could have in addressing issues of experimental discrepancies shared by *all* models.]

[JDC: Let's add a sentence that, while these are our plans, we also want to adapt to new experimental opportunities that come along, as well as respond to persistently difficult challenges by ensuring the challenge is repeated with new data as necessary.]

SAMPL6: Cyclohexane/water and octanol/water distribution coefficients. Building on the success of distribution coefficient measurements in SAMPL5 and their surprising ability to motivate rapid advances in physical modeling methodologies [refs], we will measure cyclohexane-water distribution coefficients at pH 7.4 for a new batch of commercially-available drug-like molecules for for which no data is currently available. [JDC: What about some fragment molecules as well as drug-like molecules? Maybe the set can be broken into classes

of varying chemical complexities? In general, I think clever experimental design ideas—indications we have really thought through how to maximize value—will be rewarded by the reviewers.] Given the routine nature of octanol-water distribution coefficient measurements and indications that their prediction may not be computationally tractable [**?**, 19] desite the heterogeneous structure of the wet octanol phase [**?**], we plan to also collect octanol-water distribution coefficient measurements for the same compounds. [JDC: pK_a prediction proved to be difficult but critical for SAMPL5. What about either measuring pK_as and *providing* these for the compounds in SAMPL6, or picking compounds that we are essentially certain will be free of protonation state effects (for at least a subset of compounds) so we can focus on forcefield issues.The following year, we can ask them to predict both pK_a *and* distribution coefficients.]

SAMPL7: pKa measurements for drug-like molecules. While much less complex than protein-ligand affinities, distribution coefficient measurements still conflate several issues which are still complex, namely protonation state and tautomer prediction, as well as transfer into different environments which may contain small but important quantities of cosolvents. Thus, we will likely need to turn to separating these issues to improve our handling of them one at a time. For SAMPL7, then, our tentative plan is to measure pKa values for an extensive set of drug-like molecules in water and provide data specifically on pKa values, thereby separating the issues of predicting protonation state from those of transfer. [JDC: I don't believe the Sirius T3 is capable of pK_a measurements for other solvents. I'm not even sure if that concept is meaningful since K_a is defined with regard to free proton concentrations in water.

SAMPL8: pKa measurements and distribution coefficients. In the next challenge, SAMPL8, we would recombine the pKa and transfer issues in a way to maximize learning – specifically, we will not only measure distribution coefficients but also measure pKa values for the same set of compounds, allowing participants to (a) predict distribution; (b) predict pKa; and (c) predict partitioning.

SAMPL9 and SAMPL10. Several other avenues are of interest for future datasets as well, especially for SAMPL9 and SAMPL10. New computational techniques are targeting membrane permeability [**?**, **?**], and this is experimentally accessible (see support letters from Pfizer and Merck), leading to potential interest in new datasets and challenges focused there. Alternatively, partition/distribution into other environments aside from cyclohexane or octanol may provide significant value, especially given dielectric constant issues posed by current force fields [ref]. Solubility measurements may also be suitable for a late-stage SAMPL challenge such as SAMPL10, since solubility predictions are now beginning to become tractable [**?**, **?**, **?**] (with Schrödinger also working on amorphous solubility prediction). [JDC: Let's be concrete and suggest introducing solubility for SAMPL9 and PAMPA for SAMPL10.]

Experimental plan: Experimental data will be collected in collaboration with partners in the pharmaceutical industry, roughly following the model used for SAMPL5, where the Chodera lab sent a graduate student (Bas Rustenburg) to Genentech to conduct cyclohexane-water logD measurements by adapting a high-throughput mass spectrometry workflow already in use within Genentech [**?**]. To collect physical property measurement datasets for SAMPL6-10, the Mobley and Chodera labs will send graduate students on several-week visits or internships to industry collaborators to collect targeted datasets. Working with industry collaborators (see Letters of Collaboration from Genentech, Pfizer, and Merck) gives us substantial access to equipment and high-throughput measurement workflows—such as the Sirius T3 from Sirius Analytical (which can measure partition/distribution coefficients, pK_as, and solubilities for molecules with titratable groups), automation equipment, and compound libraries—for the purposes of rapidly collecting targeted datasets. As noted, we already know that this model will work given our experience in SAMPL5 [23], and our pharma partners see the value of this data and the SAMPL challenge to the modeling community.

[JDC: I think this section should go at the beginning of this Aim.] This new experimental data is critical to make SAMPL what it ought to be. While the community has already benefitted tremendously from SAMPL, as indicated by the tremendous number of publications, the citation count, and lessons learned, the benefits are not what they could have been if the effort had funding. For example, we planned to ensure that the measured $\log D$ values in SAMPL5 spanned the full dynamic range allowed by the experiment with roughly equal populations at all values of $\log D$, but because our experimental work ended when the relevant Genentech internship did, this was not accomplished [14, 23]. [JDC: Be careful here. Let's not say "we're bad at planning experiments", but instead say "with funding, we can do more elaborate experiments".] With funding, we will be able to continue experiments work until the necessary data is collected rather than terminating it at a specific time point dictated by industry funding. This snafu had downstream consequences in that many participants observed that calculated values spanned a larger dynamic range than the experimental values, but it is unclear if this is an artifact of the data set measured (which has a relatively limited dynamic range), or because of issues with the calculations or with the experiments themselves [14, 21–23]. With more dynamic range available in the experimental data, and the ability to do follow-up

experiments (as we will have here) this would have been resolved, providing further impetus to improve models.

Thus, in Aim 1, we will generate a range of new, high quality physical property data on distribution, partition, pKa prediction, and membrane permeability to facilitate community learning via a series of new SAMPL challenges which are the focus of Aim 4 below. This will extend the prior success of SAMPL challenges on hydration and distribution between water and cyclohexane, focusing the community on problems that are of tremendous relevance to accurately predicting biomolecular interactions but without the additional complexity introduced by simulations of proteins or nucleic acids. This data thus paves the way to applications in more complex systems such as those addressed in Aims 2 and 3. [JDC: This paragraph seems unnecessary. Instead, we should focus on clearly explaining the overview at the beginning of the aim.]

Aim 2: Measure binding of novel host-guest complexes for introductory ligand binding challenges.

Physical property challenges focus on the behavior of small molecules and how this is modulated by environment, introducing a number of issues relevant to drug discovery—but binding in host-guest systems introduces a wider variety of challenges relevant to biomolecular interactions, but without the full array of challenges seen in protein-ligand interactions, as reviewed recently [24]. Thus, we believe that new data and SAMPL challenges on host-guest binding are a critical step towards accurately modeling biomolecular interactions. [JDC: This doesn't do a convincing job of communicating the distinct role host-guest challenges play. What relevant challenges are present in host-guest binding? What challenges are absent? Why is that important? How does this differ from the physical property challenges? Imagine selling someone in drug discovery on why they should care.] Already, over the past several SAMPL challenges, host-guest systems have provided key tests for modeling of binding interactions, driving a great deal of learning about challenges such as how co-solvents and protonation modulate binding and the limitations or pitfalls of different methods for calculating binding free energies [13, 15, 24, 25]. [JDC: When possible, can we avoid empty phrases like "driving a great deal of learning" and instead give the reader a coherent explanation, like "highlighted the role that protonation state effects can play in binding phenomena, resulting in errors of up to X kcal/mol if these effects are neglected"?] Host-guest binding proves remarkably difficult to model accurately [**?**], probably in part due to force field limitations—itself an important insight and one which is leading to new developments [26]. [JDC: Assuming convergence is achieved, that leaves only forcefield limitations and protonation/tautomer effects. How can these contributions be dissected or separated, and how can is failure or success specifically able to drive progress?]

drug	**features**
memantine	adamantane; 1:1
saxagliptin	adamantane; 1:1
premarin	steroid
pancuronium	steroid
varenicline	1:1 vs 1:2
valsartan	pKa 4.37
omeprazole	pKa 4.77
ranolazine	pKa 7.17; epitopes
pradaxa	pKa 3.87; epitopes
nilotinib	epitopes; pKa 6.3
sensipar	epitopes; folding
vyvance	diamine; epitopes; folding
minocycline	tetracyclin; amino aniline

Table 1. **Selected drugs whose binding to CB[n] hosts will be assayed for SAMPL6, 8, and 10 challenges (SA 2.1).** [JDC: Need a stand-alone description here. Why are they important? What is the basic idea behind the challenge? Imagine someone is just reading the figures and captions!]

Here, we design a series of SAMPL challenges focused on two general classes of host-guest systems—cucurbiturils and derivatives (SA 2.1) and deep-cavity cavitands (SA 2.2)—both of which build on studies from prior SAMPL challenges. [JDC: It would help to set these two systems in perspective in a paragraph orienting the reader before diving into the subaims. Why field two host-guest systems? How are they different, from the ten-thousand-feet perspective? What do we hope to accomplish with these challenges? What kinds of phenomena will they drive progress in modeling?]

Subaim 2.1: Cucurbituril-based receptors as model binding systems

Cucurbituril derivatives for host-guest binding. Building on previous success with cucurbit[n]uril (abbreviated CB[n]) experiments for previous SAMPL challenges [27–29], we will conduct a series of new experiments on these receptors for three new SAMPL challenges, [JDC: Just SAMPL6-8? Why only three?] with experimental work conducted by co-investigator Isaacs, an expert on these systems who has provided data for previous SAMPL challenges. CB[n] receptors are particularly well suited for the SAMPL challenges because they exhibit: (1) high binding affinities toward suitable guests in water comparable to protein-ligand affinities (routinely μM to nM; occasionally pM to fM) [30–36], (2) high selectivities between structurally related guests which translate into large $\Delta\Delta G$ values [37], (3) low molecular weights (1–2 kDa) permitting high levels of theory to be used, and (4) highly restricted conformational degrees of freedom, eliminating challenges associated with protein targets with many soft degrees of freedom and slow motions. For SAMPL6, 8, and 10, we will resynthesize a series of CB[n]-type receptors of increasing complexity, measure K_a values, and determine host-guest stoichiometry and geometry toward biologically relevant guests in order to stringently test methods for predicting binding. [JDC: What is meant by "biologically relevant guests"?] Figure 1 shows the chemical structures of three hosts—Me$_4$CB[8] [38], glycoluril

hexamer [39], and acyclic CB[n]-type receptors [40–45] which span the range from preorganized macrocyclic host to uncharged acyclic but preorganized host to highly charged acyclic host.

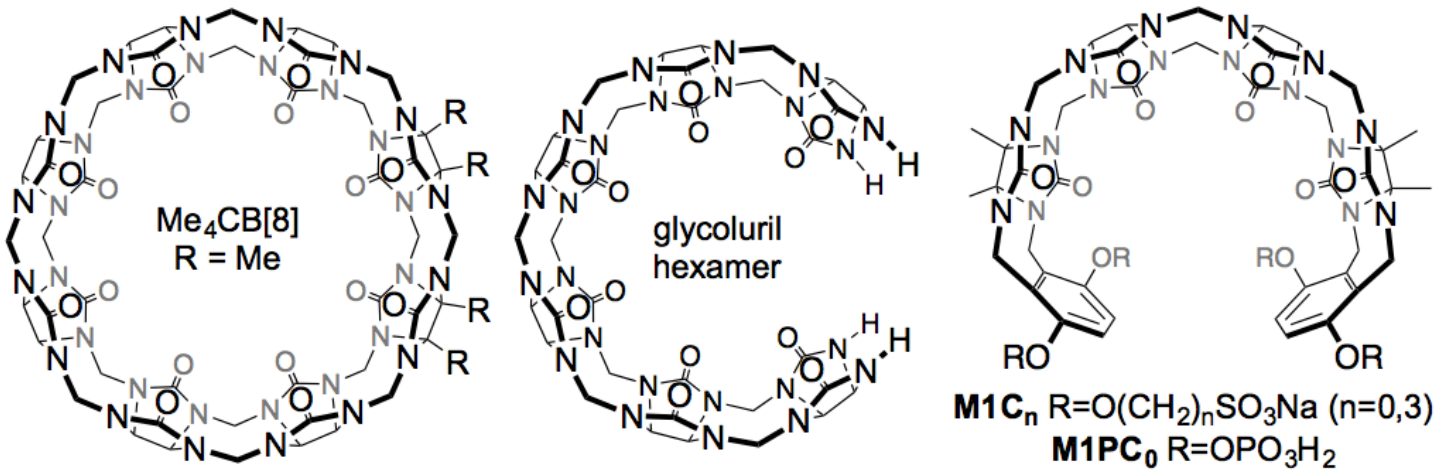

Figure 1. **Structures of Me$_4$CB[8], glycouril hexamer, and acyclic CB[n]-type receptors.** [JDC: Need a stand-alone description here so someone scanning the figures will know what is going on, which challenges they will be used for, why they are interesting. Think about someone who is only scanning the figures—the majority of the reviewers!—and make sure they get a complete story.

SAMPL6, 8, and 10 cucurbituril challenges. *For SAMPL6*, we will measure K_a and ΔH values, stoichiometry, and geometry for the interaction of Me$_4$CB[8] (a soluble CB[8] derivative) with 15 guests (selected top drugs, Table 1) by either direct or competition isothermal titration calorimetry (ITC), UV/Vis or fluorescence indicator displacement assay, or NMR competition experiments, as previously [29–31, 46]. Our selection of Me$_4$CB[8] binding to top drugs allows us to modulate the computational complexity by: 1) changing host flexibility (e.g. Me$_4$CB[8] can exhibit ellipsoidal deformation) [38], 2) allowing the possibility of binary or ternary (e.g. 1:1 and/or 1:2 host:guest) complexes [47–49], 3) using drugs with several potential binding epitopes to induce sampling issues. Host:guest stoichiometry and geometry (e.g., which binding epitope is complexed) will be addressed by ITC n values, Job plots monitored by UV/Vis or NMR [50], and by ^{1}H NMR complexation induced changes in chemical shifts [51]. All three sets of studies will be conducted in phosphate buffered saline (pH 7.4 with physiological salt) which introduces its own complexities due to salt competition for binding [24, 52]. For *SAMPL8*, we will focus on binding of the same 15 drugs (Table 1), but to glycoluril hexamer. This is an ideal next step as this host exhibits increased conformational dynamics, and influences the number and energy of solvating (and unusually coordinated) water molecules implicated in the high binding constants for CB[n]-guest complexes [36, 53]. The selected drugs include several with pK_a values in the 3.8 to 7.4 range; given that CB[n]-type receptors (like biomolecular receptors) can induce pKa shifts in their guests of up to 4 pKa units [54–56], this will test how well models can predict these effects. Additionally, it will couple nicely with the focus on pK_a values in Aim 1. *SAMPL10* will shift to acyclic CB[n]-type receptors (e.g. M1C$_3$, M1C$_0$, and M1PC$_0$ that contain anionic solubilizing groups attached via different linker lengths. As in SAMPL2, these acyclic CB[n]-type receptors introduce conformational complexity, and water interactions play a key role. Moreover, the presence of 4 anionic groups in close proximity to the cavity are expected to have a significant influence on the balance between ion-dipole interactions, and the solvation of the free host.

Subaim 2.2. Gibb deep cavity cavitands for host-guest studies

History of octa-acid SAMPL challenges. During SAMPL4 [57] and SAMPL5 [58] we focused on two hosts: the octa-acid 1 (R = H) and another octa-acid derivative with four methyl groups positioned at the portal of the binding pocket (1, R = Me). These studies used isothermal titration calorimetry (ITC) to measure the thermodynamics of (1) host **1** (R = H) complexing a range of carboxylate guests, and (2) the binding of carboxylate and trimethylammonium guests to both hosts (**1**, H = H and Me; Figure 2). In both cases ^{1}H-NMR titration was also used in a confirmatory role for ITC-derived free energies of binding. SAMPL5 emphasized how differences in the shape of the hydrophobic pocket of the host can have a profound affect on affinity for some guests.

Novel deep cavity hosts probe the effects of binding site charge constellations. For future SAMPL challenges, we will expand on the range of hosts by including **2** and **3** in our ITC studies (Figure 2). Like cavitand **1**, host **2** is an octa-acid derivative. However, the four benzoate groups are relocated from the extreme exterior in the case of **1**, to the rim of the binding pocket in **2**. We surmise that this will have a direct effect on the binding of charged guests, but more subtly, an indirect effect on guest complexation via changes to the solvation of the empty

host. Octa-trimethylammonuim cavitand ("positand" **3**) has the same overall architecture as host **1**, but inverts the charges on the water solubilizing exterior coat. While it is not yet clear if this switch in groups relatively remote from the pocket will directly affect guest complexation, results from related systems suggest it can (unpublished).

Guests for the five proposed ITC studies will be obtained from commercial sources, focusing on molecules that probe the limitations of current force-fields as well as new data as it is gathered.

SAMPL6-10 deep cavity cavitand challenges. The host-guest challenge for SAMPL6 will focus on how well the effect of host carboxylate substituent location can be predicted, and will involve hosts **1** and **2** with a set of five, previously uninvestigated guests. SAMPL7 will provide a second iteration of this experiment to test algorithmic improvements in predictive modeling following SAMPL6 by comparing hosts **1** and **3** with a different set of guests. We anticipate that because of the relative remoteness of the charged groups in these two hosts, the effects of switching charges will be subtler than the differences between **1** and **2**. SAMPL8 will consider the effect of common biologically-relevant counterions/salts salts on guest binding, comparing the effects of NaCl and NaI on the complexation of five guests to **1**. We have previously shown that iodide has a weak affinity for the binding pocket of **1**, whilst sodium ions have an affinity for the outer carboxylates [59], requiring modeling to capture the differential affinities of these ions in addition to guest affinities to successfully model the observed affinities. SAMPL9 will follow up on this by examining the effects of these same two salts on the complexation of five guests to **3**, again giving the modeling community time to incorporate algorithmic improvements following SAMPL8. While we have not yet quantified salt affinities to host **3**, we expect the iodide to have affinity for both the pocket and the positively charged solubilizing groups. For SAMPL10 we will consider the effects of co-solvents on the binding of five guests to **1** and **2** to probe the effect of co-solvent competition for the binding site, as well as effects cosolvents may have in weakening the hydrophobic effect.

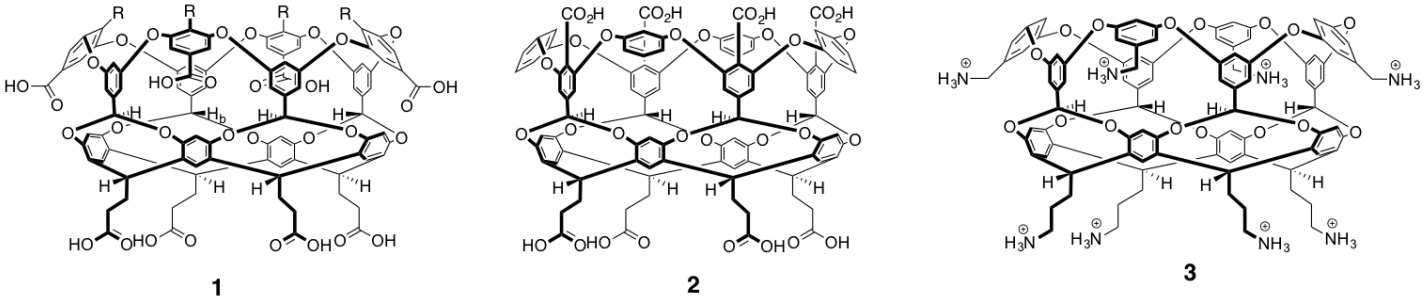

Figure 2. **Gibb deep cavity cavitands for SAMPL6-10.** [JDC: Need a stand-alone description here so someone scanning the figures will know what is going on.

Aim 3. Develop model protein:ligand systems that isolate specific modeling challenges found in more complex pharmacologically relevant systems. A major goal of our effort is to drive advances in the quantitative modeling of protein:ligand interactions. While the Drug Design Data Resource (D3R [6]) effort provides community blind challenges for biomolecular targets of pharmaceutical interest, these targets generally contain a daunting number of complexities that frustrate the ability for current methodologies to achieve quantitative accuracy, resulting in poor performance [CITE]. For example, while kinases are targets of great interest to drug discovery, blind challenges involving kinase targets conflate issues of slow protein conformational dynamics [**?**], protonation state effects of both protein [**?**] and ligand [**?**, **?**], charged ligands, and the modeling of complex divalent salt environments and phosphorylation state effects along with the standard computational challenges of conformational sampling and forcefield accuracy. While the value of these exercises as an accurate prospective benchmark of current-generation model accuracy is unquestionable, **the ability of blind challenges on complex pharmaceutical targets to rapidly advance the field of quantitative predictive modeling is limited.**

Instead, **our philosophy is to identify model protein:ligand systems with the goal of *isolating* individual accuracy-limiting effects in iterative cycles of prospective blind community challenges.** This process focuses the field on identifying and evaluating multiple solutions to the accuracy-limiting effects (such as how to deal with ligand and protein protonation-state issues [**?**], slow protein conformational dynamics, etc.) free from other complicating factors, allowing a direct evaluation of how well the phenomena of interest are modeled. Datasets collected for these blind challenges then become standard benchmark datasets for retrospectively examining the effectiveness of modeling approaches in treating these effects to facilitate comparisons of methodologies in publications, while future iterations or variations of the same SAMPL experiment allow iterative refinement and prospective blinded evaluations of methodologies. In this way, this cycle of blind challenges utilizing model systems

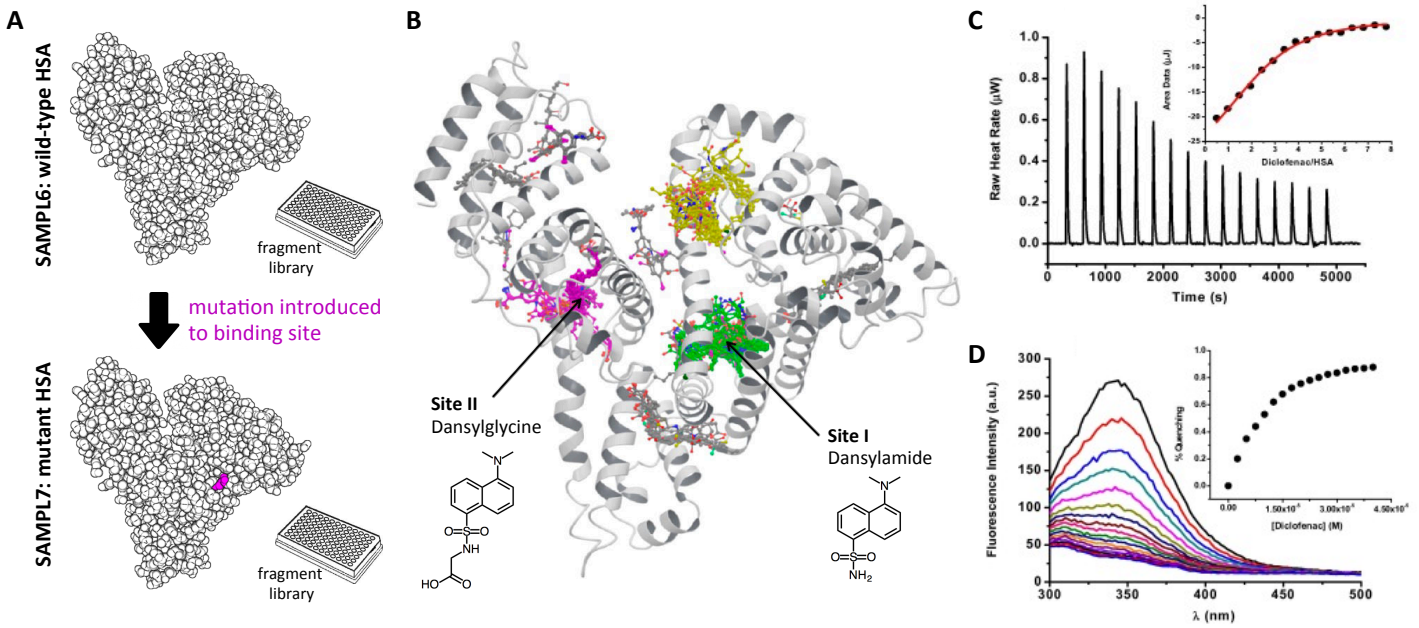

Figure 3. **The SAMPL6/7 protein:ligand challenge focuses on soluble drug fragment binding to human serum albumin (HSA).** *(A)* SAMPL6 will introduce recombinant human serum albumin (HSA) as a target, against which a library of 96 small soluble drug fragments will be assayed. By introducing a mutation in one of the binding pockets, we will create a second challenge target for SAMPL7. *(B)* HSA, the most abundant plasma protein, has at least eight known binding sites, with major well-characterized sites (green, Sudlow's Site I; purple, Sudlow's Site II) observed to bind a variety of drugs, often modulating their pharmacokinetics [**?**] (figure from [**?**]). Two fluorescent probes—dansylamide and dansylglycine—have been shown to bind with $\sim\mu$M affinity and high selectivity to Site I and Site II, respectively; both exhibit binding-enhanced fluorescence at 480 nm, and can be used to site-specifically probe ligand affinities by competition. *(C)* Binding affinities of soluble molecules can easily be measured by isothermal titration calorimetry (ITC); here, the ITC titration of HSA by diclofenac (a Site II ligand [**?**]) is shown [**?**]. *(D)* HSA tryptophan fluorescence quenching can also be used to measure ligand binding affinity; here, HSA titration by diclofenac is shown, with the inset plot showing percent quenching at 346 nm [**?**]

can rapidly drive progress in rapidly overcoming scientific hurdles limiting quantitative accuracy.

While model protein-ligand systems have a long and storied history of driving progress in individual research laboratories (such as the Shoichet T4 lysozyme mutants [**?**]), their power in blind community challenge cycles is amplified by leveraging community participation. An excellent example of this was the collection of a challenge dataset featuring the binding of small, rigid charged molecules to bovine trypsin for SAMPL3 [**?**], which rapidly focused the field on the deficiencies of current alchemical free energy methodologies in treating the binding of charged ligands. Within two years, multiple laboratories had developed and disseminated convergent practical solutions to effectively handle charged ligand binding that are now adopted as best practices [**?**, **?**].

SAMPL6-10 model protein:ligand challenges. For the SAMPL6-10 challenges, we propose to introduce a new model protein:ligand system each year, with challenges fielded for each system for at least two consecutive years to allow iterative methodology improvement and assessment. Immediately following the challenge, challenge data (including all primary data) will be published and released as a version-controlled benchmark dataset for retrospective evaluation. The first challenge (introduced in SAMPL6) will focus on modeling the binding of small soluble drug fragments to a relatively rigid protein with multiple weak binding sites, isolating the ability of current-generation modeling approaches to model weak and multiple binding effects. Because rapidly focusing the field on current challenges in predictive modeling requires the ability to adapt to deficiencies identified by D3R/SAMPL challenges of the previous year, subsequent model systems will be rapidly identified and developed using a new informatics platform we have developed to identify tractable model systems.

SAMPL6: Assessing predictive modeling to multiple weak binding sites with the binding of small soluble fragments to human serum albumin (HSA). Human serum albumin (HSA), the most abundant blood plasma protein, has the remarkable ability to bind a great variety of small molecule drugs in multiple binding sites (Figure 3B) [**?**]. As a result, HSA is not only an excellent model system for isolating the challenge of binding multiple weak ligands to a stable rigid protein, it is also a pharmacologically relevant system due to its ability to drastically modulate drug pharmacokinetics [**?**]. HSA has at least *eight* known binding sites, with numerous crystal structures available for drugs binding to two predominant sites (Sudlow Site I and II) [**?**]. Small soluble molecules

resembling drug fragments have previously been shown to have a high likelihood of detectable binding to HSA ($\geq$90% of small druglike fragments, as detected by SPR [**?**]), providing an experimentally-tractable diverse set of ligands spanning several orders of magnitude in affinity [CITE]. As current advanced methodologies such as alchemical free energy calculations currently assume a single well-defined binding site with high affinity [**?**], this dataset will allow the isolation of the effect of weak multiple binding from the majority of other counfounding factors in protein-ligand binding.

Recombinant HSA will be expressed in *E. coli* and purified via refolding from inclusion bodies [**?**], and will be defatted at low pH to ensure the resulting protein is free of the complications of both glycosylation and bound fatty acids found in plasma-isolated HSA [**?**]. Recombinant expression will also allow a mutant form of HSA (engineered via quick-change single-primer mutagenesis) to be fielded for a SAMPL7 iteration of this challenge (Figure 3B). We will obtain a diverse library of 96 soluble drug-fragment-like molecules in pre-plated format for which HSA-ligand affinities are not available in the literature as dry compound, and assay them for HSA binding using automated isothermal titration calorimetry (ITC) (Figure 3C), with the goal of characterizing the overall binding affinity of the compound to HSA. The same ligands pre-plated in DMSO format will be used to conduct a separate set of fluorescence titration assays (monitoring tryptophan fluorescence quenching) and competition assays in which the site-specific fluorescent probes daynsylamide (Site I) and dansylglycine (Site II) will be used to measure site-specific affinities for Sites I and II (Figure 3D), allowing participants to validate whether they predicted the correct binding site and, if so, the site-specific affinity.

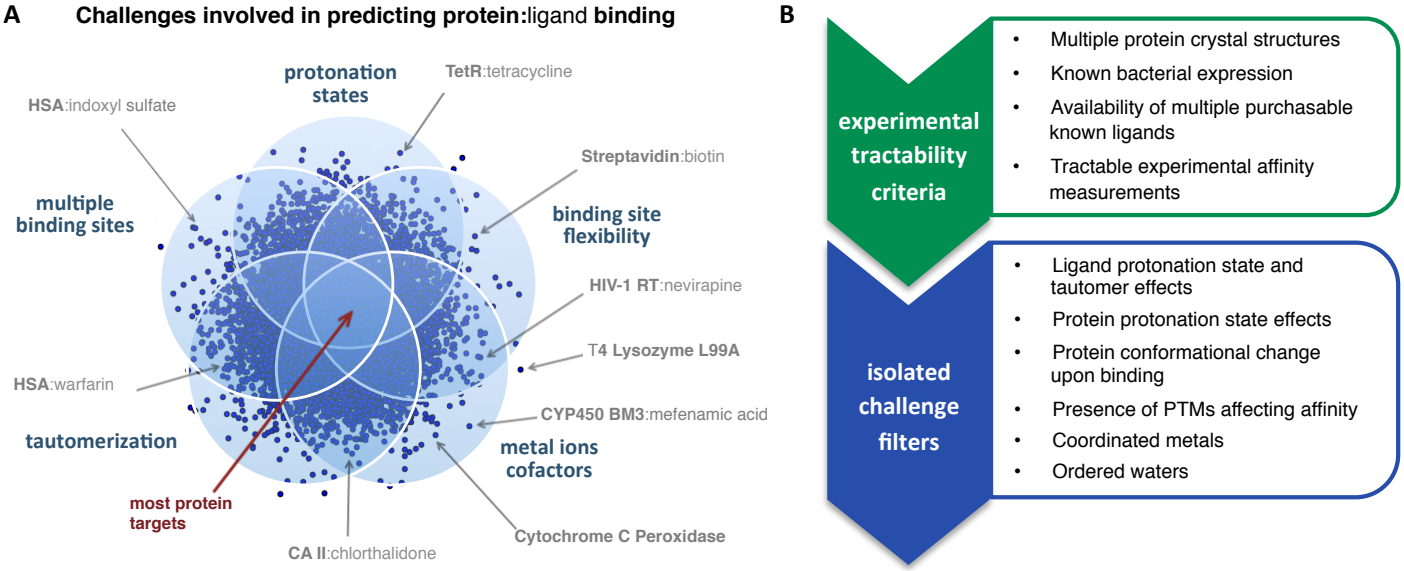

Figure 4. **Mining model protein:ligand systems to focus on individual modeling challenges via a structural and chemical informatics platform.** SAMPL7-10 will feature the introduction of new model protein:ligand systems designed to focus on individual challenges judged to be of critical immediate importance following current D3R/SAMPL blind competitions. *(A)* Since most protein targets of pharmaceutical interest feature a multitude of conflated challenges to quantitative accuracy, our goal is to identify model protein targets that isolate individual effects to focus community efforts by fielding new blind challenges. Some example protein:ligand pairs and the conceptual challenge categories they fall under are shown in the figure: T4 Lysozyme L99A [4], HSA [**?**], Carbonic anhydrase II (CAII) [**?**, **?**], Cytochrome C Peroxidase [**?**], Cytochrome P450 BM3 M11 (CYP450 BM3) [**?**], HIV-1 Reverse Transcriptase (HIV-1 RT) [**?**, **?**], Streptavidin [**?**], Tet repressor protein (TetR) [**?**, **?**]. *(B)* In order to rapidly develop new experimentally and computationally-tractable model protein:ligand systems, we have developed a structural and chemical informatics system that applies successive filters to the set of all potential protein:ligand systems for which structural data is available. Suitable model systems should meet all experimental tractability criteria (green box) and only possess a few of challenging properties, ideally only one (blue box).

SAMPL7-10: Rapid, responsive development of new model systems using a novel informatics platform. We have developed a novel informatics platform called TargetExplorer aimed at identifying new protein targets that can be rapidly developed into experimentally- and computationally-tractable model systems focusing on individual challenges. This tool—which will be made accessible to other laboratories via an easy-to-use web interface during the course of this project—successively filters all protein:ligand complexes identified in the PDB according to a list of criteria that allow facile development as a model protein:ligand binding system, as well as identification of systems that isolate individual challenges in modeling accuracy. Experimental tractability includes: (1) the availability of multiple protein:ligand crystal structures; (2) known bacterial expression (e.g. from PDB `EXPRESSION_SYSTEM` records); (3) the capacity to bind a wide dynamic range of ligands (determined via data available in ChEMBL); (4) the availability of multiple known ligands that can be purchased (via ZINC); (5)

tractability of experimental affinity measurements, such as known ligands with potentially fluorescent scaffolds (for fluorescence competition assays), highly soluble ligands (for ITC), or ligands above a minimal mass (for SPR or MST). A number of additional filters annotate potential experimentally tractable systems for suitability as a model system that isolates individual challenges, such as: charged ligands or potential ligand protonation state or tautomer [**?**] effects (deduced from predicted aqueous protonation/tautomer energies); potential protein protonation state effects (deduced from MCCE2 calculations [**?**]); protein conformational changes (deduced from variation in protein conformation or the presence of unresolved loops in protein:ligand crystal structures); the presence of post-translational modifications that may affect affinity (deduced from Uniprot annotations); coordinated metals (identified in crystal structures); ordered waters (present in multiple crystal structures); etc.

The Chodera lab has developed an automated wetlab for the purpose of rapidly developing model protein systems using bacterial expression techniques (see Equipment and Facilities). Potential targets matching desired challenge criteria will be screened for bacterial expression using high-throughput cloning, transformation, and expression testing, with purity and yield assessed by capillary electrophoresis on a Caliper GXII. Targets will be screened for stability in various buffers using Thermofluor [**?**]. Ligands identified via the informatics platform to span a wide dynamic range of binding affinities will be purchased as dry powder stocks and prepared for assay by highly accurate gravimetric solution preparation techniques using a Quantos automated balance. A wide variety of biophysical techniques are available to provide accurate, quantitative measurements of protein-ligand binding affinities, including fluorescence (if fluorescent probe ligands are available), absorption (e.g. Soret band shifts), automated isothermal titration calorimetry (provided ligands are sufficiently soluble), surface plasmon resonance, microscale thermophoresis (MST), luminescence, and alphascreen; all except MST are fully automated.

Our approach to developing challenge datasets will be twofold: First, small molecules similar to known ligands will be purchased and assayed, with the presumption that these molecules are likely to have measurable affinities. Second, site-directed mutants will be introduced to modulate the binding affinities of known ligands using single-primer quick-change mutagenesis, which can be performed and screened for expression in 96-well format. Challenge datasets will therefore consist of a matrix of protein mutants and ligands, providing a rich dataset to deeply explore the effects of interest.

Aim 4. Field iterated community blind challenges to advance quantitative biomolecular design. While Aims 1–3 focus on the critical need for targeted datasets to improve deficiencies in physical modeling of protein-ligand interactions, **the value of this data is amplified enormously in the strategic release of this data through multiple iterations of coordinated SAMPL blind challenges and related activities**. These blind challenges are designed to test the state of the art, solicit new methodological and force field innovations, allow comparative evaluation of methods, and drive downstream improvements. The new, progressive, targeted nature of the data generated for this purpose means that SAMPL challenges can now build on one another, and for success in later challenges, participants must build on lessons learned from prior challenges. SAMPL challenges will and subsequent data release activities will therefore facilitate rapid iterations of application, learning, and improvement.

SAMPL blind challenges. SAMPL challenges will have submission deadlines yearly, though not every data type or component of Aims 1-3 will be annual. The full timeline for SAMPL challenges (Figure 5) will be made available on the SAMPL website [`https://drugdesigndata.org/about/sampl`] at the outset, allowing participants to plan their work and select what challenges to be involved in. [JDC: Historically, how many participants have we had? How many do we expect? Can we communicate how many researchers this has impacted even *before* funding?] As experimental data for each component becomes available and is curated, input files and challenge details will be made available online (and advertise via e-mail lists from prior SAMPLs and CCL), at least six months prior to the challenge deadline; data not yet available at that time will be held for a subsequent challenge (with the exception of three months for year 1 due to startup timescales). As in prior SAMPLs, submissions will be handled by an automated web upload service on the SAMPL website (which will be migrated to separate hosting if the D3R effort is not renewed after its current term) which validates submissions to ensure that they meet format standards we specify along with the challenge details. As in SAMPL4 and SAMPL5, analysis will also be conducted by our automated Python framework, as will e-mail return of plots and data analysis associated with each submission. [JDC: We can likely move to a web-based framework. Even if it is just automated IPython notebook generation, it sounds be better than "e-mail return", which is pretty close to saying we will telegram or dispatch a horse-borne courier with the results.] Participants will be allowed to submit multiple sets of predictions per SAMPL component to allow for comparative assessment of models. One new ingredient will be that each participant will be required to designate one submission as their "primary" submission which will be included in the formal ranking of performance; other submissions will still be analyzed and receive an assessment of performance, but only one will be formally ranked. We find that participants learn a great deal from being able to compare multiple methods or protocols, but providing some participants with more "shots on goal" than others in the formal ranking

can be seen by some as unfair. All participant submissions and methodology descriptions will (as before) be made available publicly on the website, along with participant information (except for participants who specifically request to remain anonymous prior to submission). [JDC: Do we also make the aggregate statistics and historical performance available on the website? Can we?]

Our goal is not just to run blind challenges, but to advance modeling by helping participants identify both modeling failures and potential solutions. To achieve this, we provide guidance to participants as to what known modeling issues we expect may be relevant when providing details on each SAMPL component. For a host-guest system, for example, we might highlight known buffer/salt effects, protonation state challenges, and point out previous work on sampling challenges, with pointers to the relevant experimental work and to modeling work from past SAMPL challenges and elsewhere [24]. This helps participants design their approach. **Additionally, we will run reference calculations using standard best practices.** This serves several purposes: It provides a test of the current methods we select (usually those we view as current standard best practices) and current force fields; it helps facilitate learning—we announce what calculations we plan to perform, make input files available in a wide variety of formats [ref HG, DC overviews and Shirts], and others can repeat our calculations with a different method but same system and force field to compare methods, or swap force field but keep the method and system fixed to compare force fields, etc.; and it allows us to conduct sensitivity analysis, as by varying the conditions of our simulations (protonation state, tautomer, etc., [14]) we can see how much this impacts calculated values and thereby how important it is, even if participants don't do these tests. Reference calculations have, for example, helped us highlight the importance of a small amount of water in cyclohexane for accurately calculating log D values, determine that an incorrect tautomer could affect calculated values by many log units [14], and highlight how forcefield modifications could significantly improve results on hydration free energies [12]. To further aid follow-up studies, we will make the input files, results, and simulation workflows used for our reference calculations—along with the data—available via open software development portals such as GitHub and Docker Hub.

Physical methods are only valuable if they can reliably outperform alternate methods, so **a new focus of SAMPL6-10 will be selecting quality null models and running them to provide a point of comparison for participants**. While SAMPL efforts in the past have used some null models, lack of manpower has required that these be particularly crude (such as "guess zero" models [refs]) which provide no predictive value. In this work, we will announce our selected null models when we open each challenge component, so participants know what their results will be compared with. Statistical analysis of SAMPL performance, and comparison with nulls, will continue to be an important component of our work [refs].

Following submission and analysis of each SAMPL challenge, challenge results will be released and discussed, with SAMPL workshops allowing more formal presentations on and discussion of results in years 1, 3, and 5. Workshops will run every two years at the request of past participants, and will be co-run and co-hosted with D3R Grand Challenge workshops (see support letter). During the off years, SAMPL challenges will still run, but discussion of and dissemination of results will be via asynchronous means (as discussed below) and a "virtual workshop" consisting of talks and interaction over Google Hangouts or YouTube Live (formerly Hangouts On Air). While coordination with D3R will mostly be at the level of workshops, we will also work with them ensure that SAMPL challenge submission deadlines are offset from deadlines for D3R to allow maximum community participation in both efforts.

Dissemination of results and data. Rapid dissemination of results is critical so that new insights can be rapidly spread to the community. We will continue to publish special issues of the *Journal of Computer Aided Molecular Design* (JCAMD) collecting publications related to each year's SAMPL challenges (see Letter of Support). To ensure immediate availability of reports, we will strongly encourage prepublication sharing of results and analysis, including both slides and posters from SAMPL meetings (via F1000 Research) and paper preprints (via bioRxiv). We also want to ensure that participants are able to learn from one another by exchanging ideas outside of formal workshops and meetings. While this has happened in the past—for example, when participants using similar methods work together after the SAMPL meeting to identify the origin of these discrepancies [refs]—we hope to accelerate this kind of collaboration. To facilitate more open communication between the community, we will use collaboration software—such as Slack, which facilitates scientific communication for the NASA/JPL Mars Rover teams and NSF antarctic scientific research teams—to build a community discussion platform, allowing interested individuals to openly exchange ideas, tests, and work together to learn from their work.

Each dataset will have a life cycle of collection, curation, blind challenges, and public dissemination. In the past, the unfunded nature of SAMPL has forced us to primarily emphasize the *blind challenges* and pre-challenge *curation* aspects, with isolated forays into collection [ref SAMPL5, ref Peat trypsin]. This work considers the full life cycle for the first time, with Aims 1–3 dealing primarily with collection and pre-challenge curation.

Post-challenge, datasets will receive additional curation and then be released to the community as standard test or benchmark sets that allow retrospective evaluation of methodologies on high-quality data [ref benchmark set paper]. The FreeSolv dataset, for example, includes a large number of calculated and experimental hydration free energies from SAMPL0–4, and now represents a standard benchmark dataset for hydration free energy calculations containing the majority of available high-quality data [FreeSolv and SAMPL0-4 hydration refs]. Post-challenge curation will receive new attention here; in the past, lack of resources has always prevented follow-up experimental work, even when the data clearly indicated it was warranted (such as puzzling issues with dynamic range for logD values in SAMPL5 [ref]). The requested budget will ensure that these experiments will now be conducted, allowing computation to drive collection of additional experimental data—potentially by different methodologies—when warranted. Dissemination is the final stage in the data life cycle (see Resource Sharing Plan for details); our driving philosophy is to make the data (including primary data, processed data, and our analysis of challenge submissions) available freely and publicly with permanent, citeable DOIs; ensure relevant data is also deposited in standard community repositories (e.g. BindingDB [REF]); and guarantee data longevity via backup hosting on library archival facilities (such as UCI's DASH [link]).

Because prior SAMPL work focused mainly on challenges themselves, dissemination of the data has never been a major focus. As part of this project, we will retrieve old SAMPL experimental data, submissions, and analysis from where it is archived and make it available along with the new data generated here.

We will also push for containerization of tools and methods in conjunction with other efforts such as AutoDesk's Molecular Design Toolkit and the NSF Molecular Sciences (MolSci) initiative. Our vision is that long-term, instead of participants submitting a set of predictions, they would also submit the entire workflow they applied via Docker containers in a way which would allow others to reproduce their work if needed, or to repurpose the workflow for new applications, ensuring not just that SAMPL *results* get disseminated, but that the *methods* and *workflows* themselves spread.

TIMELINE

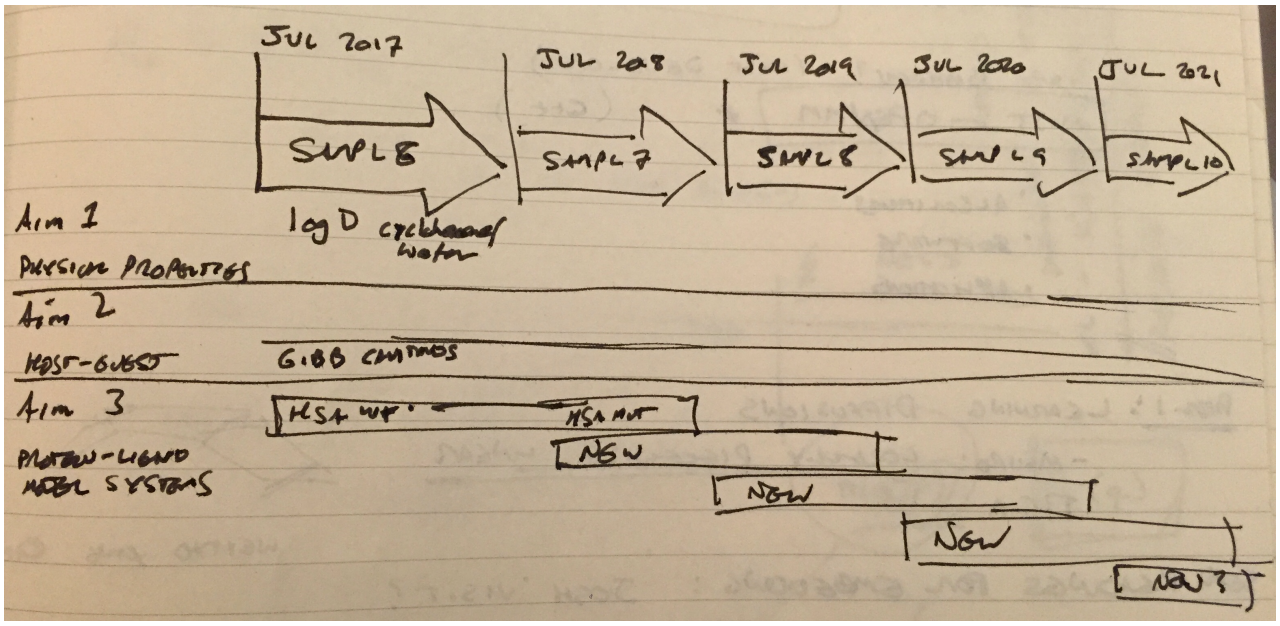

Figure 5. **Timeline for SAMPL activities.** [Insert a figure to summarize SAMPL iterations and coordinated challenges in each category (Aim 1, Aim 2, Aim 3).]

[JDC: I think we need to cut this section and stick to a few sentences of commentary on the figure.]

Yearly challenges will drive progress. [DLM: This has a lot of overlap with Aim 4 in terms of what timing details it gives, so I'll have to shrink one or the other to remove redundancy.] Our series of SAMPL challenges is progressive in difficulty, with subsequent challenges building on those from previous years and driving progress towards biomolecular binding for pharmaceutical datasets as covered by D3R. Each year's SAMPL challenge involves several components with staggered deadlines set to maximize scientific benefit. Tentative deadlines and the full project Timeline (Figure X) will be made available as soon as the project kicks off in order to allow workers to plan. Inputs will be available on our website at least 6 months prior to the challenge deadline (except in the first year, when data collection will require a more limited window of 3 months). Automated analysis of

submissions – which are required to be in a standard format – will already be online prior to the submission deadline, so immediately following the deadline, participants will receive automated reports giving the experimental results, a detailed summary of the statistics for each of their submissions and how they compare to other methods. This will also include distribution of the method descriptions that participants submit along with their predictions, allowing challenge participants to see how top performing methods differed from their own approaches. Results from our reference calculations will also be distributed at the same time. At this time participants will also receive information directing them to Slack channels (see above) for discussion of the challenge and results, as well as publication information.

Follow-up information exchange facilitates improvements. Immediately following return of challenge results, follow up communication on Slack channels focused on each challenge component will spur participants to share what they did, what they learned from it, and begin learning from one another and from us and our reference calculations. Normally, challenge participants have explored different aspects of the challenge, with some looking at details of convergence, others comparing force fields and water models, and so on. Often, this information only comes out only in publication, long after the challenge deadline and any meetings which might result. Our vision is to use the more rapid conversation that comes via Slack to propagate this information earlier. Individual messages in Slack have permanent links, meaning that participants who share here will have a permanent record of their contribution to the science – reducing concerns that ideas shared pre-publication might not receive credit. This will be the primary means of communication regarding SAMPL for eight weeks after the challenge deadline. After this eight week period, a virtual meeting will wrap up the particular challenge component and allow participants to begin refocusing their efforts on the next challenge, with ten months until the next deadline for that component and four months until the data becomes available. This will provide enough time to incorporate lessons learned. Overall meetings – virtual (years 2 and 4) or in-person (the end of years 1, 3, and 5) will wrap up the overall challenge; these will be timed to follow the last component challenge of each of the relevant years.

Publication reaches a broader audience. During the initial eight weeks after the submission deadline/return of results, participants will be encouraged to begin preparing their results for publication. We highly encourage the use of preprints to get material out to the broader audience as quickly as possible. However, as noted, we will continue to work with JCAMD to have regular special issues focused on SAMPL. Deadlines for these will be yearly, and the review/revision process is normally expected to take around two months from submission until when papers begin appearing online. Composition of the special issue itself has to take longer, as the issue can't be finalized until all papers are typeset. Thus our priority, for rapid dissemination is on more rapid means of communication, though the permanent academic record of the work will continue to need to be traditional publishing.

Datasets generate lasting value as benchmarks. As discussed in Aim 4, all SAMPL data will be made easily and publicly available to serve the community in the future as retrospective tests, and will be deposited in standard repositories such as BindingDB when applicable. In some cases, the cumulative value of data will be profound. For example, some of the host-guest challenges do not incorporate a particularly large number of compounds, but over the lifespan of SAMPL6-10, the amount of data accumulated will be considerable. Selected SAMPL data sets are expected to become especially valuable as standard benchmark sets and will receive additional attention by being singled out by the community as such [24]. In these ways, SAMPL will have a lasting effect on the modeling community.

COLLABORATION MANAGEMENT PLAN

We have a strong previous history of successful collaboration, with Mobley and Chodera having co-authored roughly a dozen publications since 2006, as well as several workshops and other initiatives. Mobley, Isaacs, and Gibb have also worked together to coordinate past SAMPL challenges, and Mobley and Gibb a previous NSF workshop. PI Mobley will oversee the entire project, with Mobley and Chodera working together on Aim 1 (Mobley working with west coast pharma; Chodera with east coast), Isaacs responsible for Aim 2.1, Gibb for Aim 2.2, and Chodera for Aim 3. Mobley and Chodera will conduct aim 4, involving the other co-investigators as needed. Meetings will consist of a Google Hangout monthly and an in-purpose planning meeting once yearly which will take place at the SAMPL workshops during the years those are offered, and a separate meeting doing the alternate two years. Chodera and Mobley will communicate more frequently due to the interlinked nature of their work. Publications are expected to be largely dictated by the overall Timeline noted above, with an experimental publication associated with each challenge component being prepared for distribution to participants along with their results. Conflict resolution is expected to be straightforward given the extensive past history of interaction and collaboration between the investigators, but if any serious difficulties arise, Michael Gilson (UCSD) will be consulted as an arbiter, given the need this initiative has for close connections with D3R.

OUTLOOK

Physical methods have been slow to achieve their promise in binding prediction, in part because truly significant innovations are so hard to recognize due to a lack of standard tests and benchmarks, and in part because of an “applications first” approach which seems to plague our community where we rush to apply our methods to problems of pharmaceutical relevance without ensuring they can tackle simpler, better-understood problems first. Here, we propose an innovative extension of the successful series of SAMPL blind challenges, generating novel experimental data to drive improvement of the methods in our field and help them become pharmaceutically relevant – beginning with relatively simple physical property prediction and progressing to challenging problems in biomolecular recognition via a series of carefully designed intermediate steps. SAMPL already has a strong track record of success, and funding will ensure an even greater impact on the community and that it is around as a valuable resource for years to come. The proposed series of carefully tailored challenges will focus our community on a variety of problems which we *can* realistically resolve in the near term, resulting in dramatic improvements in the field’s ability to do computational molecular design.

Bibliography and References Cited

[1] Mobley, D. L. and Klimovich, P. V.: Perspective: Alchemical free energy calculations for drug discovery. J. Chem. Phys. 137(23): 230901, January 2012.

[2] Deng, N., Forli, S., He, P., Perryman, A., Wickstrom, L., Vijayan, R. S. K., Tiefenbrunn, T., Stout, D., Gallicchio, E., Olson, A. J., and Levy, R. M.: Distinguishing Binders from False Positives by Free Energy Calculations: Fragment Screening Against the Flap Site of HIV Protease. J. Phys. Chem. B. 119(3): 976–988, January 2015.

[3] Lim, N. M., Wang, L., Abel, R., and Mobley, D. L.: Sensitivity in binding free energies due to protein reorganization. Journal of Chemical Theory and Computation. July 2016.

[4] Mobley, D. L., Graves, A. P., Chodera, J. D., McReynolds, A. C., Shoichet, B. K., and Dill, K. A.: Predicting absolute ligand binding free energies to a simple model site. J. Mol. Biol. 371(4): 1118–1134, August 2007.

[5] Boyce, S. E., Mobley, D. L., Rocklin, G. J., Graves, A. P., Dill, K. A., and Shoichet, B. K.: Predicting ligand binding affinity with alchemical free energy methods in a polar model binding site. J. Mol. Biol. 394(4): 747–763, December 2009.

[6] Gathiaka, S., Liu, S., Chiu, M., Yang, H., Stuckey, J. A., Kang, Y. N., Delproposto, J., Dunbar, J. B., Carlson, H. A., Burley, S., Walters, W., Amaro, R. E., Feher, V., and Gilson, M. K.: D3R Grand Challenge 2015: Evaluation of Protein-Ligand Pose and Affinity Prediction. J Comput Aided Mol Des. 2016.

[7] Nicholls, A., Mobley, D. L., Guthrie, J. P., Chodera, J. D., Bayly, C. I., Cooper, M. D., and Pande, V. S.: Predicting Small-Molecule Solvation Free Energies: An Informal Blind Test for Computational Chemistry. J. Med. Chem. 51(4): 769–779, February 2008.

[8] Nicholls, A., Wlodek, S., and Grant, J. A.: The SAMP1 Solvation Challenge: Further Lessons Regarding the Pitfalls of Parametrization. J. Phys. Chem. B. 113(14): 4521–4532, April 2009.

[9] Mobley, D. L., Bayly, C. I., Cooper, M. D., and Dill, K. A.: Predictions of Hydration Free Energies from All-Atom Molecular Dynamics Simulations. J Phys Chem B. 113: 4533–4537, January 2009.

[10] Geballe, M. T., Skillman, A. G., Nicholls, A., Guthrie, J. P., and Taylor, P. J.: The SAMPL2 blind prediction challenge: Introduction and overview. J Comput Aided Mol Des. 24(4): 259–279, May 2010.

[11] Geballe, M. T. and Guthrie, J. P.: The SAMPL3 blind prediction challenge: Transfer energy overview. J Comput Aided Mol Des. 26(5): 489–496, April 2012.

[12] Mobley, D. L., Wymer, K. L., Lim, N. M., and Guthrie, J. P.: Blind prediction of solvation free energies from the SAMPL4 challenge. J Comput Aided Mol Des. 28(3): 135–150, March 2014.

[13] Muddana, H. S., Fenley, A. T., Mobley, D. L., and Gilson, M. K.: The SAMPL4 host–guest blind prediction challenge: An overview. J Comput Aided Mol Des. 28(4): 305–317, March 2014.

[14] Bannan, C. C., Burley, K. H., Chiu, M., Shirts, M. R., Gilson, M. K., and Mobley, D. L.: Blind prediction of cyclohexane-water distribution coefficients from the SAMPL5 challenge. 2016.

[15] Yin, J., Henriksen, N. M., Slochower, D. R., Chiu, M. W., Mobley, D. L., and Gilson, M. K.: Overview of the SAMPL5 Host-Guest Challenge: Are We Doing Better? J Comput Aided Mol Des. 2016.

[16] Ignjatović, M. M., Caldararu, O., Dong, G., Muñoz-Gutierrez, C., Adasme-Carreño, F., and Ryde, U.: Binding-affinity predictions of HSP90 in the D3R Grand Challenge 2015 with docking, MM/GBSA, QM/MM, and free-energy simulations. J Comput Aided Mol Des. pp 1–24, August 2016.

[17] Deng, N., Flynn, W. F., Xia, J., Vijayan, R. S. K., Zhang, B., He, P., Mentes, A., Gallicchio, E., and Levy, R. M.: Large scale free energy calculations for blind predictions of protein–ligand binding: The D3R Grand Challenge 2015. J Comput Aided Mol Des. pp 1–9, August 2016.

[18] Sunseri, J., Ragoza, M., Collins, J., and Koes, D. R.: A D3R prospective evaluation of machine learning for protein-ligand scoring. J Comput Aided Mol Des. pp 1–11, September 2016.

[19] Bannan, C. C., Calabró, G., Kyu, D. Y., and Mobley, D. L.: Calculating Partition Coefficients of Small Molecules in Octanol/Water and Cyclohexane/Water. Journal of Chemical Theory and Computation. 12(8): 4015–4024, August 2016.

[20] Luchko, T., Blinov, N., Limon, G. C., Joyce, K. P., and Kovalenko, A.: SAMPL5: 3D-RISM partition coefficient calculations with partial molar volume corrections and solute conformational sampling. J Comput Aided Mol Des. pp 1–13, September 2016.

[21] Paranahewage, S. S., Gierhart, C. S., and Fennell, C. J.: Predicting water-to-cyclohexane partitioning of the SAMPL5 molecules using dielectric balancing of force fields. J Comput Aided Mol Des. pp 1–7, August 2016.

[22] Klamt, A., Eckert, F., Reinisch, J., and Wichmann, K.: Prediction of cyclohexane-water distribution coefficients with COSMO-RS on the SAMPL5 data set. J Comput Aided Mol Des. pp 1–9, July 2016.

[23] Rustenburg, A. S., Dancer, J., Lin, B., Feng, J. A., Ortwine, D. F., Mobley, D. L., and Chodera, J. D.: Measuring experimental cyclohexane-water distribution coefficients for the SAMPL5 challenge. bioRxiv. 063081 pp, July

2016.
[24] Mobley, D. L. and Gilson, M. K.: Predicting binding free energies: Frontiers and benchmarks. bioRxiv. 074625 pp, September 2016.
[25] Bhakat, S. and Söderhjelm, P.: Resolving the problem of trapped water in binding cavities: Prediction of host-guest binding free energies in the SAMPL5 challenge by funnel metadynamics. J Comput Aided Mol Des. 2016.
[26] Yin, J., Fenley, A. T., Henriksen, N. M., and Gilson, M. K.: Toward Improved Force-Field Accuracy through Sensitivity Analysis of Host-Guest Binding Thermodynamics. The Journal of Physical Chemistry B. 119(32): 10145–10155, August 2015.
[27] Ma, D., Glassenberg, R., Ghosh, S., Zavalij, P. Y., and Isaacs, L.: Acyclic cucurbituril congener binds to local anaesthetics. Supramolecular Chemistry. 24(5): 325–332, May 2012.
[28] Cao, L. and Isaacs, L.: Absolute and relative binding affinity of cucurbit[7]uril towards a series of cationic guests. Supramolecular Chemistry. 26(3-4): 251–258, March 2014.
[29] She, N., Moncelet, D., Gilberg, L., Lu, X., Sindelar, V., Briken, V., and Isaacs, L.: Glycoluril-Derived Molecular Clips are Potent and Selective Receptors for Cationic Dyes in Water. Chem. Eur. J. pp n/a–n/a, August 2016.
[30] Cao, L., Šekutor, M., Zavalij, P. Y., Mlinarić-Majerski, K., Glaser, R., and Isaacs, L.: Cucurbit[7]uril·Guest Pair with an Attomolar Dissociation Constant. Angew. Chem. Int. Ed. 53(4): 988–993, January 2014.
[31] Liu, S., Ruspic, C., Mukhopadhyay, P., Chakrabarti, S., Zavalij, P. Y., and Isaacs, L.: The Cucurbit[n]uril Family: Prime Components for Self-Sorting Systems. Journal of the American Chemical Society. 127(45): 15959–15967, November 2005.
[32] Mock, W. L. and Shih, N. Y.: Structure and selectivity in host-guest complexes of cucurbituril. The Journal of Organic Chemistry. 51(23): 4440–4446, November 1986.
[33] Assaf, K. I. and Nau, W. M.: Cucurbiturils: From synthesis to high-affinity binding and catalysis. Chem Soc Rev. 44(2): 394–418, January 2015.
[34] Moghaddam, S., Yang, C., Rekharsky, M., Ko, Y. H., Kim, K., Inoue, Y., and Gilson, M. K.: New Ultrahigh Affinity Host-Guest Complexes of Cucurbit[7]uril with Bicyclo[2.2.2]octane and Adamantane Guests: Thermodynamic Analysis and Evaluation of M2 Affinity Calculations. Journal of the American Chemical Society. 133(10): 3570–3581, March 2011.
[35] Shetty, D., Khedkar, J. K., Park, K. M., and Kim, K.: Can we beat the biotin–avidin pair?: cucurbit[7]uril-based ultrahigh affinity host–guest complexes and their applications. Chem. Soc. Rev. 44(23): 8747–8761, 2015.
[36] Biedermann, F., Uzunova, V. D., Scherman, O. A., Nau, W. M., and De Simone, A.: Release of High-Energy Water as an Essential Driving Force for the High-Affinity Binding of Cucurbit[n]urils. J. Am. Chem. Soc. 134(37): 15318–15323, September 2012.
[37] Isaacs, L.: Stimuli Responsive Systems Constructed Using Cucurbit[n]uril-Type Molecular Containers. Acc. Chem. Res. 47(7): 2052–2062, July 2014.
[38] Vinciguerra, B., Zavalij, P. Y., and Isaacs, L.: Synthesis and Recognition Properties of Cucurbit[8]uril Derivatives. Org. Lett. 17(20): 5068–5071, October 2015.
[39] Lucas, D., Minami, T., Iannuzzi, G., Cao, L., Wittenberg, J. B., Anzenbacher, P., and Isaacs, L.: Templated Synthesis of Glycoluril Hexamer and Monofunctionalized Cucurbit[6]uril Derivatives. J. Am. Chem. Soc. 133(44): 17966–17976, November 2011.
[40] Ma, D., Zhang, B., Hoffmann, U., Sundrup, M. G., Eikermann, M., and Isaacs, L.: Acyclic Cucurbit[n]uril-Type Molecular Containers Bind Neuromuscular Blocking Agents In Vitro and Reverse Neuromuscular Block In Vivo. Angew. Chem. Int. Ed. 51(45): 11358–11362, November 2012.
[41] Ma, D., Hettiarachchi, G., Nguyen, D., Zhang, B., Wittenberg, J. B., Zavalij, P. Y., Briken, V., and Isaacs, L.: Acyclic cucurbit[n]uril molecular containers enhance the solubility and bioactivity of poorly soluble pharmaceuticals. Nat Chem. 4(6): 503–510, June 2012.
[42] Zhang, B. and Isaacs, L.: Acyclic Cucurbit[n]uril-type Molecular Containers: Influence of Aromatic Walls on their Function as Solubilizing Excipients for Insoluble Drugs. J. Med. Chem. 57(22): 9554–9563, November 2014.
[43] Gilberg, L., Zhang, B., Zavalij, P. Y., Sindelar, V., and Isaacs, L.: Acyclic cucurbit[n]uril-type molecular containers: Influence of glycoluril oligomer length on their function as solubilizing agents. Org. Biomol. Chem. 13(13): 4041–4050, 2015.
[44] Sigwalt, D., Moncelet, D., Falcinelli, S., Mandadapu, V., Zavalij, P. Y., Day, A., Briken, V., and Isaacs, L.: Acyclic Cucurbit[n]uril-Type Molecular Containers: Influence of Linker Length on Their Function as Solubilizing Agents. ChemMedChem. 11(9): 980–989, May 2016.
[45] Zhang, B., Zavalij, P. Y., and Isaacs, L.: Acyclic CB[n]-type molecular containers: Effect of solubilizing group on their function as solubilizing excipients. Org. Biomol. Chem. 12(15): 2413–2422, 2014.

[46] Ma, D., Zavalij, P. Y., and Isaacs, L.: Acyclic Cucurbit[n]uril Congeners Are High Affinity Hosts. J. Org. Chem. 75(14): 4786–4795, July 2010.
[47] Ko, Y. H., Kim, E., Hwang, I., and Kim, K.: Supramolecular assemblies built with host-stabilized charge-transfer interactions. Chem. Commun. (13): 1305–1315, 2007.
[48] Barrow, S. J., Kasera, S., Rowland, M. J., del Barrio, J., and Scherman, O. A.: Cucurbituril-Based Molecular Recognition. Chem. Rev. 115(22): 12320–12406, November 2015.
[49] Urbach, A. R. and Ramalingam, V.: Molecular Recognition of Amino Acids, Peptides, and Proteins by Cucurbit[n]uril Receptors. Isr. J. Chem. 51(5-6): 664–678, May 2011.
[50] Connors, K. A.: Binding Constants. New York, NY, John Wiley & Sons, 1987.
[51] Masson, E., Ling, X., Joseph, R., Kyeremeh-Mensah, L., and Lu, X.: Cucurbituril chemistry: A tale of supramolecular success. RSC Adv. 2(4): 1213–1247, 2012.
[52] Márquez, C., Hudgins, R. R., and Nau, W. M.: Mechanism of Host-Guest Complexation by Cucurbituril. J. Am. Chem. Soc. 126(18): 5806–5816, May 2004.
[53] Biedermann, F., Nau, W. M., and Schneider, H.-J.: The Hydrophobic Effect Revisited—Studies with Supramolecular Complexes Imply High-Energy Water as a Noncovalent Driving Force. Angew. Chem. Int. Ed. 53(42): 11158–11171, October 2014.
[54] 'il Saleh, N., Koner, A., and Nau, W.: Activation and Stabilization of Drugs by Supramolecular pKa Shifts: Drug-Delivery Applications Tailored for Cucurbiturils. Angewandte Chemie. 120(29): 5478–5481, July 2008.
[55] Nau, W. M., Florea, M., and Assaf, K. I.: Deep Inside Cucurbiturils: Physical Properties and Volumes of their Inner Cavity Determine the Hydrophobic Driving Force for Host–Guest Complexation. Isr. J. Chem. 51(5-6): 559–577, May 2011.
[56] Ghosh, I. and Nau, W. M.: The strategic use of supramolecular pKa shifts to enhance the bioavailability of drugs. Advanced Drug Delivery Reviews. 64(9): 764–783, June 2012.
[57] Gibb, C. L. D. and Gibb, B. C.: Binding of cyclic carboxylates to octa-acid deep-cavity cavitand. J Comput Aided Mol Des. 28(4): 319–325, November 2013.
[58] Sullivan, M. R., Sokkalingam, P., Nguyen, T., Donahue, J. P., and Gibb, B. C.: Binding of carboxylate and trimethylammonium salts to octa-acid and TEMOA deep-cavity cavitands. J Comput Aided Mol Des. pp 1–8, July 2016.
[59] Carnegie, R. S., Gibb, C. L. D., and Gibb, B. C.: Anion Complexation and The Hofmeister Effect. Angew. Chem. 126(43): 11682–11684, October 2014.